From human to humanoid ears: on building an automated sibilant detector

Joachim Kokkelmans (UniVR) joachimkokkelmans@gmail.com

1. Introduction and aims

As compared to human abilities, the strength of computers lies amongst others in their higher calculation speed, mathematical precision, and the ability to repeat the same task perhaps 1.678 times in a row without making any error. Automated computational methods are thus superior in speed and accuracy to manual human work, for they are not subject to restrictions such as fatigue, distraction or vague approximations.

However, human ears are still better at distinguishing speech sounds from each other, by segmenting in their mental representation different phonemes in a continuous utterance such as e.g. [sɪbɪlənt]. To this purpose, humans are better trained to isolate relevant sounds from background noise, by relying on various acoustic cues and lexical knowledge. Computers have thus a higher processing efficacy, whereas they lack the "practical intuition" humans have.

This study aims at combining the strongest aspects of both computers and human ears, i.e. to facilitate for phonetic research the analysis and extraction of specific sounds as recognised by the computer, which can then be verified and annotated in the fastest and easiest possible way by researchers. One aims thus not only at building a fricative detector (as in e.g. Ali & al. (2001), Spinu & Lilley (2016) or Vydana & Vuppala (2016)), but also at optimising the concrete extraction and tagging process by comparing the efficacy of manual and automated methods. In order to test and quantify the efficacy of four different methods, which differ from each other in the extent to which the computer is involved, a "sibilant extraction competition" is organised, in which the author of this abstract competes against his own computer.

2. Methods

The goal of the computational method is to facilitate the researcher's task as far as simply needing her/him to upload a sound file into a program and to recuperate the extracted sibilants with their calculated characteristics (e.g. centre of gravity, skewness etc.), in the case of the fully automated method. The full procedure takes place as follows:

1. A *Praat* script (Boersma & Weenink 2005) segments the audio file in segments of 10 milliseconds each, and calculates for each one the overall intensity in dB, the centre of gravity (COG), skewness and kurtosis.
2. Adjacent series of fragments with COG, skewness and kurtosis values indicating very probable sibilantness are understood by the script as being a continuous sibilant, and their entire spectrogram is automatically extracted as one file. Silence (which can easily have a high COG) is eliminated by testing the intensity of the sound (the minimal threshold defined is 32 dB).
3. For each extracted sibilant spectrogram, a PHP script creates a graphic representation of the sound (showing mean intensity in function of frequency), accompanied by the corresponding statistical information (COG, skewness, kurtosis, duration etc.).

4. The researcher only needs, in a user-friendly interface, to corroborate perceptually the judgement of the computer and optionally to tag the occurrence of /s/ according to own criteria (e.g. position in the word, preceding vowel etc.). A .csv file is automatically generated which contains all the relevant data and can be used for exact calculations.

3. Experiment

To test and quantify which method is the most effective, an audio file containing an informant interview in Swedish (which has the sounds [s̻], [ɕ] and [s̢]) is analysed four times:

- The *fully automated* method: The *Praat* script extracts all sibilants itself, and no corrections are made by the researcher. The spectrograms are then analysed in the PHP script, and the researcher annotates each sibilant without correcting anything.
- The *half-automated* method: *Idem*, but corrections are made by the researcher in the PHP interface, and the corrected sibilants are then reanalysed.
- The *half-manual* method: The researcher segments the sibilants in the PHP interface, creating a list of times, and annotates them. A *Praat* script then extracts the spectrograms at the defined times, and the spectrograms are analysed by the PHP script.
- The *fully manual* method: The researcher extracts all sibilants manually in *Praat*, lets the PHP script analyse them, and annotates the spectrogram in the interface.

The different methods are evaluated by multiplying the time needed for the entire procedure by the amount of errors made (for the fully automated method), or by measuring the time needed.

4. Results

Preliminary experimentation seems to indicate that the combination of computer analysis and human supervision is the most effective method, showing higher processing speed than the human method on the one hand, and a lower error rate than the computer method on the other hand. A question which is still to be answered is whether better results are attained when the computer or the user segments the sibilant extracts (i.e. with the half-automated or the half-manual method).

5. References

Ali, Ahmed. M. A., Mueller, Paul & Van der Spiegel, Jan. 2001. “Acoustic-phonetic features for the automatic classification of fricatives”. In: *The Journal of the Acoustical Society of America* 109(5): 2217–2235.

Boersma, Paul & Weenink, David. 2005. Praat: doing phonetics by computer [computer program]. http://www.praat.org/.

Spinu, Laura & Lilley, Jason. 2016. “A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives”. In: *Journal of Phonetics* 57: 40-58.

Vydana, Hari Krishna & Vuppala, Anil Kumar. 2016. “Detection of fricatives using S-transform”. In: *The Journal of the Acoustical Society of America* 140(5): 3896-3907.